J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 8, Number 7 · April 2010

Automated Scoring of an Interactive Geometry Item: A Proof-of-Concept

Jessica Masters

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 8, Number 7

Automated Scoring of an Interactive Geometry Item: A Proof-of-Concept

Jessica Masters

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Masters, J. (2010). Automated Scoring of an Interactive Geometry Item: A Proof-of-Concept. *Journal of Technology, Learning, and Assessment, 8*(7). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

An online interactive geometry item was developed to explore students' abilities to create prototypical and "tilted" rectangles out of line segments. The item was administered to 1,002 students. The responses to the item were hand-coded as correct, incorrect, or incorrect with possible evidence of a misconception. A variation of the nearest neighbor algorithm was used to automatically predict one of these categories for each of the student responses. The predicted category was compared to the hand-coded category. The algorithm accurately predicted the category for 94.6% of responses.

J·T·L·A

Automated Scoring of an Interactive Geometry Item: A Proof-of-Concept

Jessica Masters
Boston College

Introduction

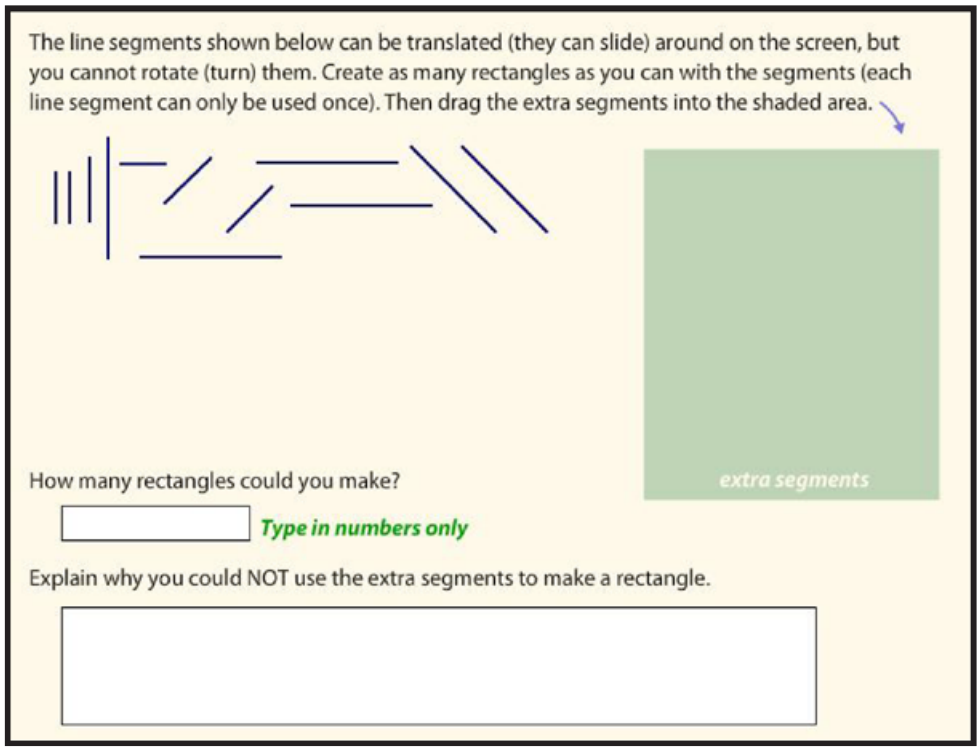
Formative assessment is an integral component of the instructional process, and classroom instruction should be tailored based on information provided by formative assessments. Popham defined two critical characteristics of formative assessment: 1) formative assessment results should be available to teachers during instruction so that the results can inform changes to instructional approaches; and 2) changes to instruction based on assessment results should be made "with the intent of better meeting the needs of the students assessed" (2006, p. 4). The simplest approach to fulfilling the first criteria of returning timely feedback is through assessments composed of closed-response items, such as traditional multiple choice items. While these types of items allow assessment results to be immediately considered when modifying instructional approaches, the depth of information provided is often limited. Open-response or constructed-response items typically provide richer information, but the scoring of these items is time-intensive, and, thus, often results cannot be instantly integrated into classroom practice. In recent years, efforts have been made to leverage technological advancements to automatically score open-response items, or items of a more interactive nature.

The bulk of these research efforts have been in the domain of automated essay scoring, but other efforts have also been made in automated scoring of open-response mathematics items. Concurrently, innovative and interactive technologies have emerged, such as Geometer's Sketchpad, allowing students to virtually interact with mathematical concepts, specifically in the area of geometry. The research presented in this paper seeks to combine interactive technology with automated scoring, in the context of geometry in the middle grades. This paper is a proof-of-concept that an interactive geometry item can be automatically scored to return instant feedback to teachers, thus informing instruction.

The interactive item described in this paper was developed as part of the Diagnostic Geometry Assessment (DGA) project, which applies findings from cognitive research on mathematical misconceptions to create online, formative, diagnostic assessments that classroom teachers can use to identify and address student misconceptions. The DGA is based on the previous successes of the similar Diagnostic Algebra Assessment project (Russell, O'Dwyer & Miranda, 2009), uses traditional closed-response items, and focuses on three geometric misconceptions. As part of the DGA development process, open-response items were created and administered to a large sample of students. The purpose of collecting these written responses was to develop closed-response distracter options representative of the correct and flawed understandings exhibited. A subset of the open-response items were online, interactive items.

The interactive items were initially developed for the same purpose as the paper-based open-response items, and were intended solely to collect student responses on which to base closed-response options. However, after the items were developed and administered, it became clear that the interactive nature of the items had great potential to engage students, elicit underlying student thought processes, and allow teachers to review students' interactions, thus allowing teachers to explore students' thought processes. While this capability alone has great potential, the items would provide even richer feedback if they were able to be automatically scored. If this were possible, a set of items could provide instant feedback, and also allow teachers to further explore the thought processes of individual students. To this end, the current paper describes the creation of an automated scoring algorithm for one of these interactive items, as a proof-of-concept that it is *possible* to automatically score this type of interactive item. A screenshot of the item discussed in this paper is displayed in Figure 1. In this item, students can move line segments to form shapes and explain their reasoning about the shapes they create.

Figure 1: Screen Shot of the Interactive Item



Related Work

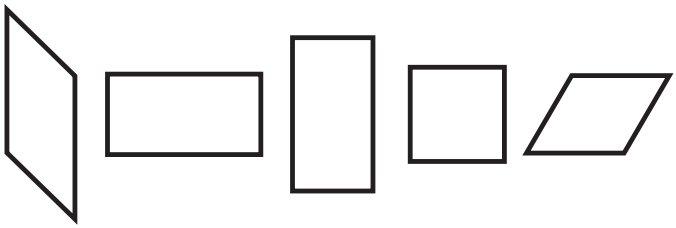
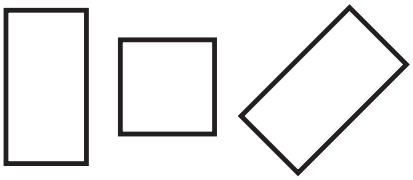
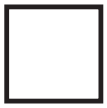
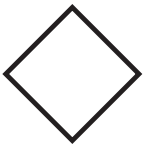
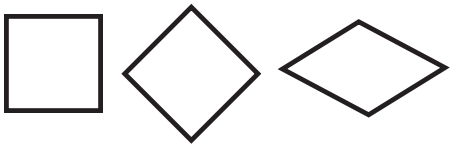
The current work relies on related research in two fields. The interactive item described in this paper was designed based on a cognitive framework describing students' understanding of geometric concepts. The automatic scoring of the item is grounded in research in the area of automated scoring.

The DGA and the Shape Properties Misconception

The DGA focuses on three misconceptions, encompassing both flawed and underdeveloped reasoning. The misconception targeted with the interactive item discussed in this paper is referred to as the Shape Properties misconception. The difficulties that students experience related to this misconception stem from students having developed a concept *image* without a concept *definition*. As Vinner and Hershkowitz describe, a concept image is a mental image that students have of a shape, without a specified definition of the shape or its properties (1980). Students who have this deficit of understanding can often identify examples of shapes, but will fail to iden-

tify examples of shapes that are not identical to their own mental image of the shape or the shape prototype, i.e., the figure does not "look like" the shape. Characteristics such as orientation and proportion affect whether students recognize certain shapes, despite the fact that these characteristics are irrelevant to the defining properties of a shape (Carroll, 1998; Clements, 2003; Clements & Battista, 1992). Table 1 shows examples of possible shape prototypes, as well as samples of figures that might *not* be recognized as the shape by a student operating under this misconception. For example, the prototypical parallelogram is slanted towards the right, and the top and bottom sides are longer than the left and right sides. Thus, a student operating under this misconception might not recognize a figure with right angles, or equal sides, as a parallelogram, even if the figure is a four-sided figure with parallel opposite sides (the definition of a parallelogram). Rather than rely solely on mental or visual images, students should have both a set of visual examples and a description of the necessary and sufficient properties that define a shape (Hershkowitz, 1989). Instruction can help students move towards this understanding.

Table 1: Sample Shapes and Prototypes

Shape	Shape Prototype	Figures Possibly not Recognized when Reasoning with a Concept Image and not a Concept Definition
Parallelogram		
Rectangle		
Square		
Rhombus		

In the context of the DGA, the term *misconception* is used broadly to refer to any systematic source of difficulty experienced by students. Thus, the Shape Properties misconception described here can be contextualized as underdeveloped reasoning, displayed by students who first understand shapes as visual images, and then progress towards deeper understandings. The misconception can also be contextualized as a flawed understanding by students who have received instruction intended to develop concept definitions, and yet their understanding persists at the level of concept images. In either case, new knowledge and instruction related to shapes (e.g., congruence, similarity, etc.) will be best integrated when it aligns with existing networks of knowledge; thus students operating with flawed or underdeveloped understanding will likely create their own interpretations and connections of new information, which can lead to systematic errors (Hiebert & Carpenter, 1992; Mestre, 1987; Resnick, 1983). If teachers are aware of misconceptions and can recognize patterns of systematic errors, they can use this information to design targeted instruction to support students in refining and reorganizing their knowledge structures (Resnick et al., 1989; Smith, diSessa, & Roschelle, 1993).

This view of evaluating and addressing misconceptions is the underlying foundation of the DGA, which seeks to provide a formative assessment that identifies students operating under misconceptions (including the Shape Properties misconception). The DGA also contains lesson plans and activities targeting the misconceptions, providing an integrated system of diagnostic feedback and instructional resources. The DGA classifies students as operating under a misconception based on the results of a group of items, not on a single item. Thus, if the interactive item described here were able to be automatically scored, it could be combined with a group of other items targeting the Shape Properties misconception, and a student's response to this set of items would be considered when providing diagnostic feedback to teachers.

The interactive item that is the focus of this paper (displayed in Figure 1, page 6) seeks to explore the Shape Properties misconception by allowing students to create rectangles out of existing line segments. Some line segments allow students to form the prototypical rectangle (top and bottom sides longer than the left and right sides, top and bottom sides aligned with the bottom of the paper/screen), while others segments can form a "tilted" rectangle, oriented at an angle. The motivation for creating this item was that a student operating under the Shape Properties misconception would likely *not* create the tilted rectangle.

It should be noted that the DGA is still in the development stages, and thus has not yet undergone validity testing or psychometric analysis. It is not the purpose of the current paper to validate that the interactive

item appropriately measures the targeted construct, or explore how the interactive item operates within a larger set of closed-response items. The current paper only seeks to determine whether the interactive item can be automatically scored using the existing data collected through the DGA project. The description of the DGA is included here to provide the context under which the interactive item was developed and the data collected. If automatic scoring is feasible on this and other interactive items, future research will explore the implications of integrating these types of interactive items with traditional closed-response items into a formative assessment framework.

Automated Scoring

The DGA project is grounded in the belief that student assessment is a central component of the instructional process, and classroom instruction should be tailored and individualized based on information provided by formative assessments (Airasian, 1991; Black & Wiliam, 1998a, 1998b; Popham, 1995, 2006, 2008). To be effective at informing instruction, assessments must return timely feedback to teachers. Thus, research efforts have been made to automatically score items that provide richer feedback, such as open-response items.

The majority of research in automatic scoring has been in the domain of automated essay scoring (AES), as evidenced by commercial products such as Project Essay Grade, IntelliMetric, the Intelligent Essay Assessor, and e-rater. AES systems have been designed for use in classroom assessments and high-stakes standardized assessments (Shermis & Burstein, 2003; Dikli, 2006). Studies evaluating the accuracy of AES systems have found generally high agreement between automated and human scoring, often higher than the agreement between inter-human scoring. Critics remain concerned about the formulaic approaches of these systems, the emphasis placed on essay constructs rather content (e.g., length rather than meaning), and the vulnerability of the systems to cheating. Further, automatic scoring of essays requires an existing cache of essays on which to train an algorithm, which is problematic for individual essay prompts designed in the classroom (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Burstein & Chodorow, 1999; Dikli, 2006; Nichols, 2005; Page, 2003; Page & Petersen, 1995; Rudner, Garcia, & Welch, 2006; Rudner, & Gagne, 2001; Rudner & Liang, 2002; Valenti, Neri, & Cucchiarelli, 2003; Wang & Brown, 2007).

A smaller base of research has explored automatic scoring of open-response mathematics items. The primary focus of this research has been on dichotomously and diagnostically assessing *mathematical expressions*, e.g., formulas, algebraic expressions, and rational expressions, mainly

at the undergraduate and graduate level (Bennett, Morley, & Quardt, 2000b; Bennett & Sebrechts, 1994; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997; Chan & Yeung, 2000a, 2000b; Sangwin, 2004; Sebrechts, Bennett, & Rock, 1991; Singley & Bennett, 1998). Bennett et al. presented the following example of an item requiring a mathematical expression as a response: "If 12 eggs cost x cents and 20 slices of bacon cost y cents, what is the cost in cents of 2 eggs and 4 slices of bacon?" (2000b, p.298). Computer-based interfaces allow the respondent to input only allowable combinations of symbols. Once the expression is entered, the scoring algorithm compares the response to the correct expression. Because expressions have an infinite number of equivalent forms, evaluating equivalency is a key component of these algorithms (Bennett, Steffen, Singley, Morley, & Jacquemin, 1997; Bennett, Morley, & Quardt, 2000b). Algorithms that automatically score mathematical expressions have been highly successful when compared to human scoring. Bennett et al. found an accuracy rate of 99.62% when comparing automated scoring to human scoring of mathematical expressions on a graduate admissions exam (1997).

More recent research has expanded the focus to the automatic scoring of quantitative and qualitative graphs, where respondents enter points, lines, and curves onto a graph (Bennett, Morley, & Quardt, 2000a). Bennett et al. presents the following two examples of items requiring a graphical response: "Suppose you are driving at a constant speed from New York to Washington, D.C., about 200 miles away. About 80 miles from New York you pass through Philadelphia, Pennsylvania. On the grid below, plot your distance from Philadelphia as a function of time from the beginning to the end of the trip." and "On the grid below, draw a triangle that has an area equal to 6 units." (2000b, p. 304, ©Education Testing Service). One automated scoring method developed for this type of item had a 98% agreement rate with human scorers when evaluated in the context of graduate admissions; further, the discrepancies in agreement were due to errors in the human, rather than the automated, scoring (ibid). This research provides a basis that automated scoring of open-response mathematics items is feasible. The current work seeks to expand this base by focusing on the subject area of geometry (beyond graphical modeling) and by focusing on a younger population (sixth, seventh, and eighth grade students).

Methods

As part of the DGA development process, paper-based open-response items and online interactive items, including the item in Figure 1 (page 6), were administered to students. The purpose of the item administration was to collect student responses for developing closed-response distracters for the DGA. The research described in this paper used this extant data to develop and evaluate an automated scoring approach to one interactive item.

Participants

There were 1,002 student responses collected to the interactive item. These students were in the classes of 24 participating teachers. Fifty-six percent of students were in eighth grade, 32% in seventh grade, and 11% in sixth grade. The teachers volunteered to participate in the DGA development process, and thus were not recruited to mirror, and are not assumed to be, a representative sample of the general teaching population.

Ninety-six percent of teachers identified themselves as White, and 4% of teachers identified themselves as Asian. Sixty-eight percent of students identified themselves as White, 8% as Black/African American, 7% as American Indian/Alaskan Native, and 4% as Asian. Sixteen percent of students identified themselves as Hispanic/Latino[1]. Seventy-one percent of participating teachers were female; 49% of participating students were female.

The majority of participating teachers (54%) resided in the South, with some teachers residing in each of the other regions (21% in the Northeast, 13% in the Midwest, and 13% in the West).[2] The large percentage of teachers from the South is likely a result of the researchers' past experience with projects and teachers from this region. The largest percentage of participating teachers taught in town locales (37%), with some teachers teaching at other locales (26% suburban, 26% rural, and 11% city).[3]

Seventy-five percent of participating teachers had been teaching for six years or more, while 25% had been teaching for five years or less. When asked about teaching mathematics specifically, 67% of teachers had taught for six years or more, 33% for five years or less. Forty-six percent of teachers had been teaching at their current school for six years or more, 54% for five years or less. Sixty-seven percent of teachers held a Masters degree as their highest level of education, while 29% held a Bachelors degree. Ninety-two percent of teachers held a regular or standard state certificate, 4% held a provisional certificate, and 4% held no certification.

Thirty-eight percent of teachers taught at schools that received Title I funding. Teachers also self-reported the percentage of their students

that received free/reduced lunch. The largest percentage of teachers (33%) responded that 26-50% of their students received free/reduced lunch; 30% of teachers responded that 0-25% of students received free/reduced lunch, 25% of teachers responded that 76-100% of students received free/reduced lunch, and 13% responded that 51-75% of students received free/reduced lunch. Eighty-nine percent of students responded that they spoke English as their first language. Finally, 92% of students responded that they had a computer in their home, and 94% of students responded that they used a computer generally either every day or a couple of times a week. Students reported less frequent use in school: 69% responded that they used a computer in school every day or a couple of times a week, 25% responded a couple of times a year.

Procedure

The responses to the interactive item were independently hand-coded by two researchers. An automated algorithm to predict codes was developed and tested, using the hand-coded scores as a comparison for the predicted scores.

Coding Student Responses

A coding rubric was created for each of the items administered for the DGA development, including the interactive item described here. Each rubric attempted to distinguish responses that were representative of a student with a strong understanding of the construct, a *knower*, a student possibly operating under the targeted misconception, a *misconceiver*, or a student who displayed neither evidence of understanding nor of misconception, a *mistaker*. To be classified as a *knower*, a student had to provide a correct and thorough response. For the interactive item, a correct and thorough response was defined as a response satisfying three criteria: 1) at least one prototypical rectangle was created, 2) the non-prototypical or "tilted" rectangle was created, and 3) the number of rectangles the student reported creating matched the number of rectangles actually created.[4] Figure 2 shows two sample student responses that were coded as *knowers*. Figure 2(a) (next page) shows the most common student response, in which students created the prototypical rectangle, the tilted rectangle, and indicated that they created two rectangles. Figure 2(b) (page 14) shows another common response, where students created the tilted rectangle, and then used the other line segments to create more than one prototypically-oriented rectangle. In responses of this type, the grading rubric allowed for flexibility as to the number of rectangles created. Technically, in Figure 2(b), there are four rectangles (one tilted rectangle and three rectangles created by the other line segments). However, because the purpose of the item was to explore students' understanding

of the properties of rectangles, specifically as those properties relate to the orientation of the rectangle, whether or not students were able to count all prototypical rectangles was considered irrelevant to this understanding. Therefore, responses of three *or* four rectangles were considered correct in responses similar to Figure 2(b) (next page). Additionally, researchers coding the items were instructed to use individual judgment when determining whether a rectangle was created or not, similar to judgment that would be used for coding paper-based drawings. For example, two of the line segments in Figure 2(b) do not line up exactly. When line segments are not perfectly aligned, or when they overlap, it could be a result of a deficit of student understanding about rectangles, or a result of human error in using the test-taking interface. Researchers subjectively interpreted this and other similar errors when determining whether to code a response as a *knower*.

Figure 2: Sample *Knower* Responses

Figure 2(a):

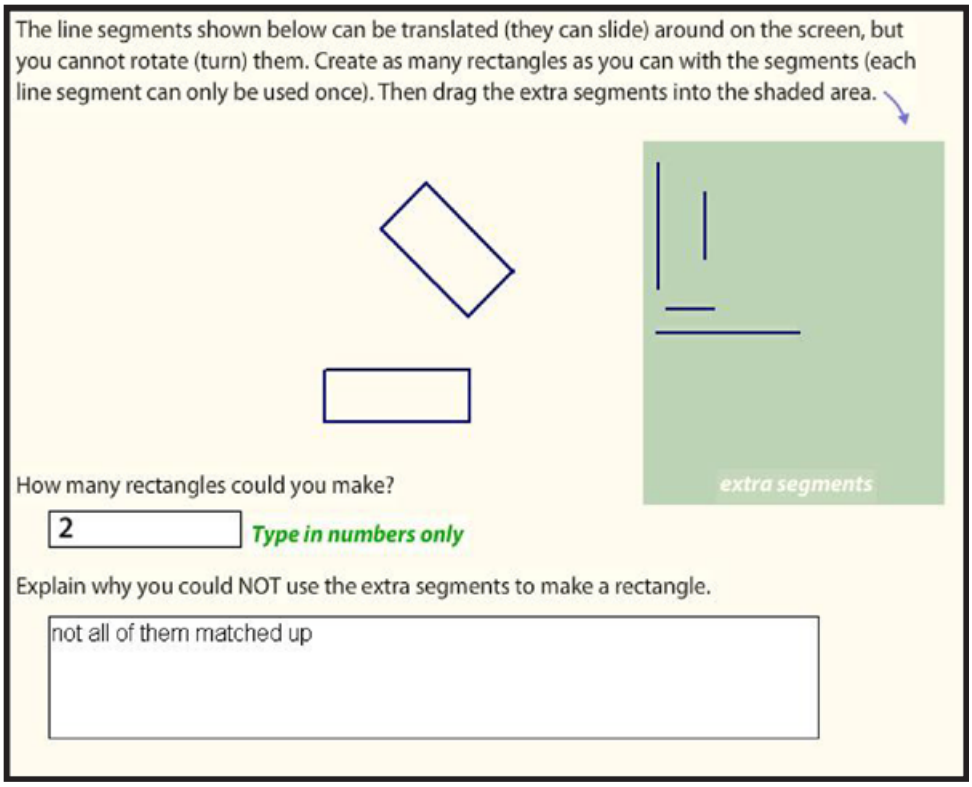
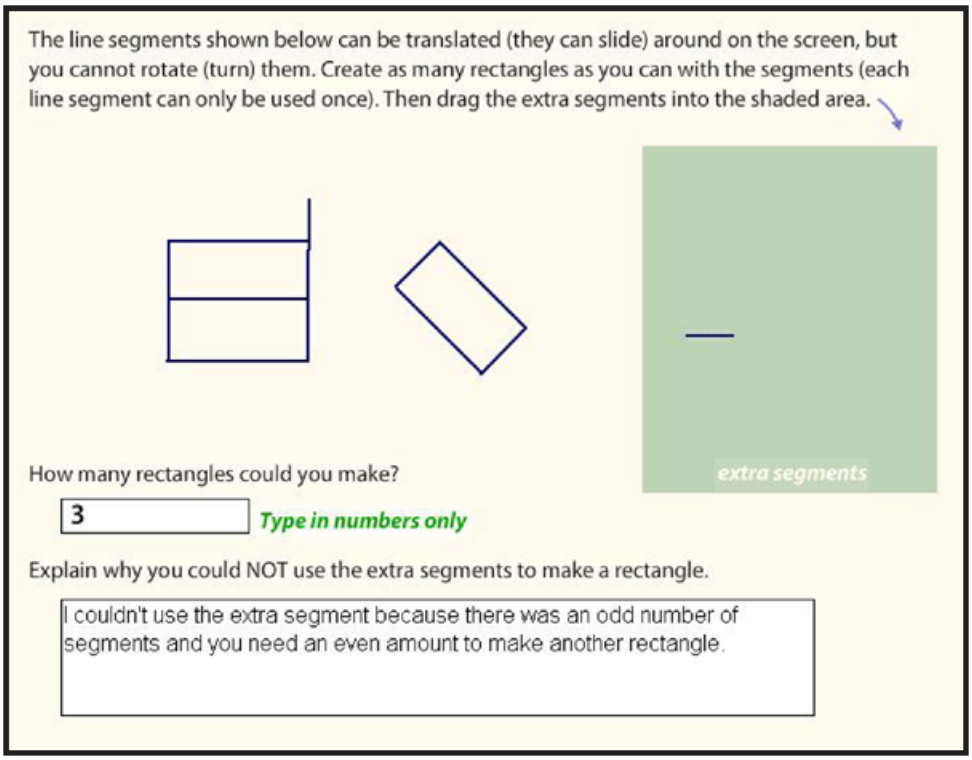


Figure 2(b):



A response was classified as a *misconceiver* if at least one prototypical rectangle was created but the "tilted" rectangle was not created. Figure 3 shows three sample responses that were coded as *misconceivers*. In Figures 3(a) and (b) (next page), the students created at least one prototypical rectangle, and responded that they created the correct number of rectangles. However, the students did not create the tilted rectangle, and thus there is some evidence that they might be operating under the Shape Properties misconception, i.e., that they only recognize rectangles that "look like" the prototypical rectangle, and not rectangles with different orientations.[5] A student response was also coded as a *misconceiver* if the tilted rectangle was created, but the number of rectangles did not include the tilted rectangle, thus indicating that the student did not, in fact, think the shape created was a rectangle. Figure 3(c) (page 16) shows a sample *misconceiver* response of this type.

Figure 3: Sample *Misconceiver* Responses

Figure 3(a):

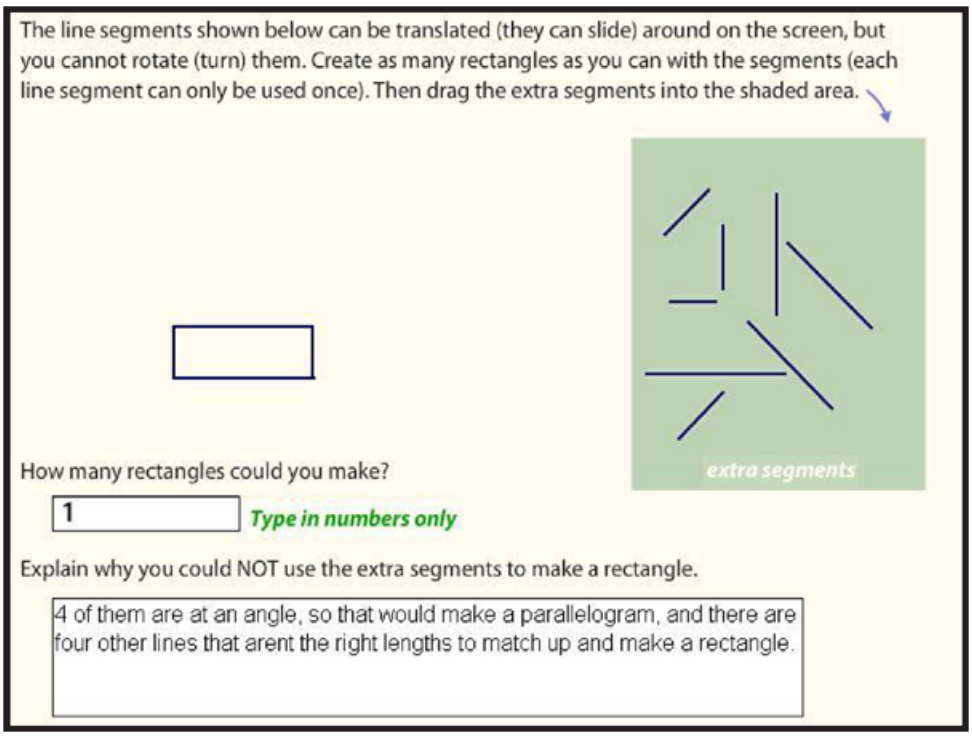


Figure 3(b):

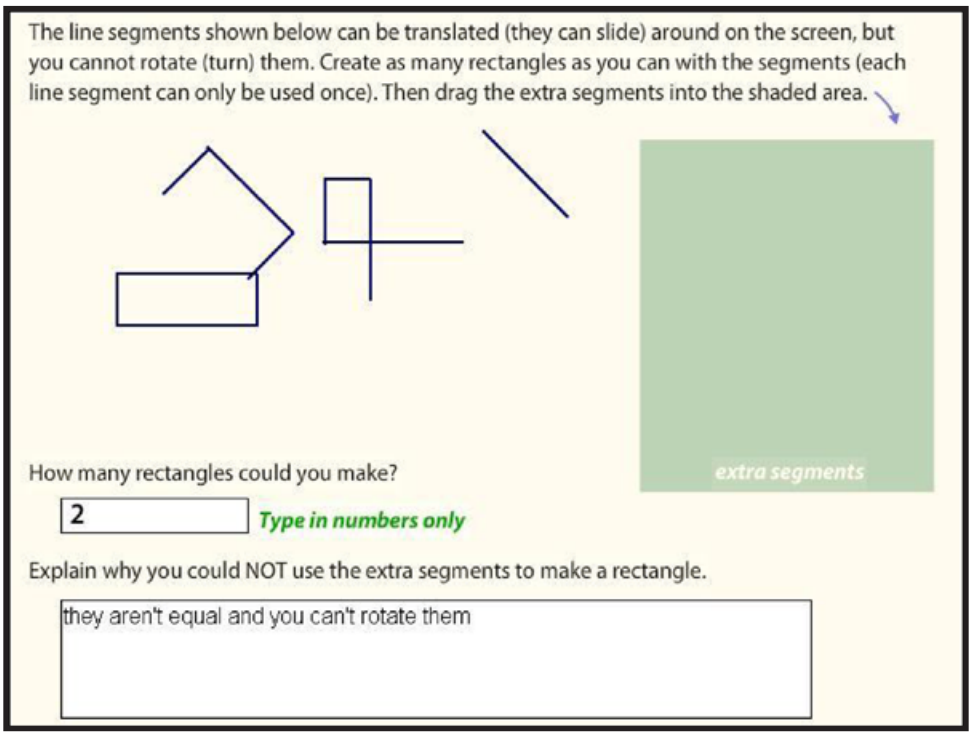
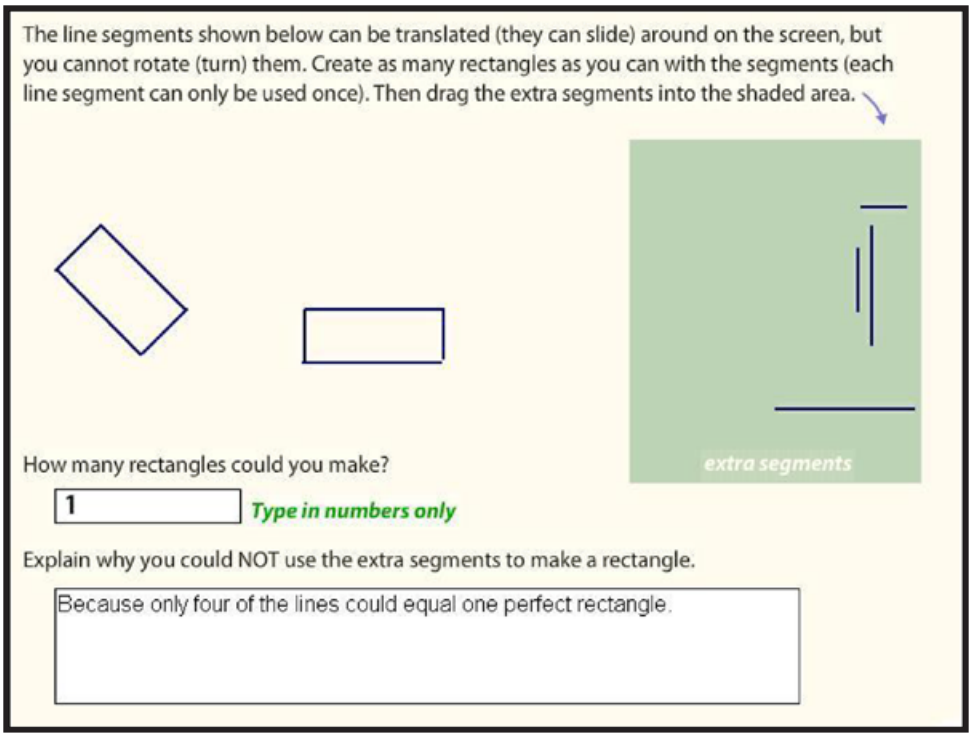
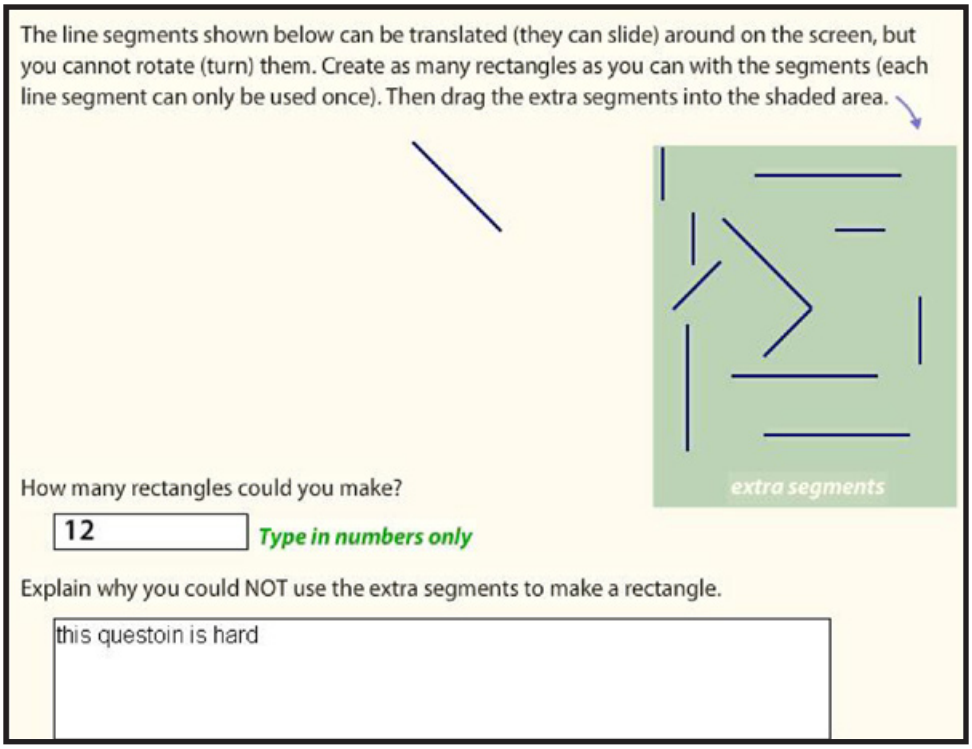


Figure 3(c):


The line segments shown below can be translated (they can slide) around on the screen, but you cannot rotate (turn) them. Create as many rectangles as you can with the segments (each line segment can only be used once). Then drag the extra segments into the shaded area.
extra segments
How many rectangles could you make?
1
Type in numbers only
Explain why you could NOT use the extra segments to make a rectangle.
Because only four of the lines could equal one perfect rectangle.

A response was classified as a *mistaker* if it did not meet the criteria of either a knower or a misconceiver. Figure 4 shows a sample response coded as a *mistaker*.

Figure 4: Sample Mistaker Response



Two independent researchers coded each response to this item. They agreed on the coding of 98.2% of the responses (Cohen's $\kappa = 0.969$). There were 15 blank responses that were discarded, since these responses can easily be coded as blank by the test delivery system. Thus, there were 969 responses on which the researchers agreed, and had been double coded as *knower*, *misconceiver*, or *mistaker*. Of these responses, 59% were coded as *knowers*, 26% were coded as *mistakers*, and 15% were coded as *misconceivers*. There were 18 responses about which the researchers disagreed.

Developing and Evaluating the Automated Scoring Algorithm

An algorithm was created to automatically predict a code for any response to the interactive item. The algorithm used a classic prediction approach from artificial intelligence, *nearest neighbors*. Prediction algorithms generally predict a category for an object based on the characteristics of that object, and on the characteristics and known categories of a separate group of objects. The *nearest neighbors* algorithm is a memory-

based prediction algorithm, meaning that it makes a prediction for a single object based on the entire database of objects for which it has both characteristics and categories. This type of memory-based algorithm is traditionally simple to implement and successful at prediction. However, these algorithms can also be expensive in time and space, and they act as a "black box," providing no explanation or reasoning behind the predictions (Breese, Heckerman, & Kadie, 1998).

The *nearest neighbors* algorithm has three basic steps. The first step is to calculate a distance between each object in the known sample, referred to as the *training set*, and the object for which a category is being predicted, referred to as the *test object*. Typically, there are multiple characteristics that describe an object, and an algorithm can consider all or some of these characteristics equally, or give more or less weight to certain characteristics, when determining how "far apart" (or similar) two objects are. The second step is to find the *neighborhood* of the test object. The algorithm is often referred to as the *k-nearest neighbors* algorithm because a number is chosen for the size of the neighborhood *(k)*. The final step in the algorithm is to predict a category for the *test object* based on the categories of the neighborhood.

As an example, consider a system that automatically filters resumes based on three keywords: New Media, Internet, and e-Commerce. Resumes are to be filtered into two categories: Interview or Reject. A human administrator first reviews five resumes and decides whether these resumes deserve an interview, or should be rejected. The categories of these five resumes, as well as the information about whether those resumes contain the keywords, form the *training set* for the automated algorithm. This information is displayed in Table 2 (next page). The algorithm must now automatically predict a category for a new resume, Resume6. The algorithm knows the characteristics of this resume- namely, whether or not it contains the three keywords. The first step for the algorithm is to find the distance between the *test object* (Resume6) and each object within the *training set* (Resume1-Resume5). For this example, we use a simple distance function ($\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$). The distances are displayed in Table 2 (next page). The next step is to determine the *neighborhood* of the *test object*. In this example, the algorithm will use a value of k=3, meaning that the neighborhood will consist of the closest three objects from the *training set*: Resume1, Resume3, and Resume5. The final step is to predict a category. Because two of the resumes in the *neighborhood* have the category "Interview," and only one has the category "Reject," the algorithm will predict that Resume6 should be classified as "Interview."

Table 2: Example *Nearest-Neighbor* Data

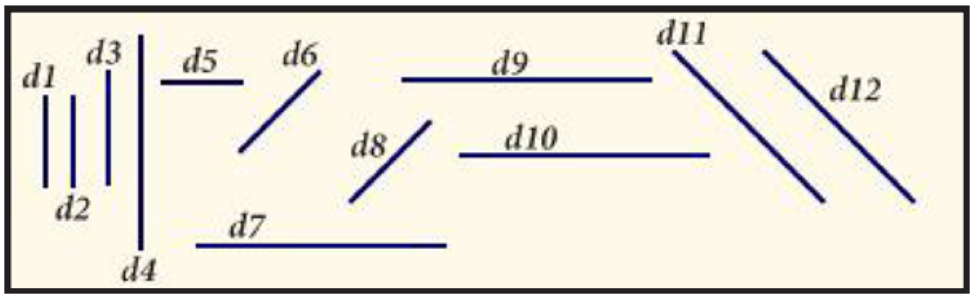
		Category	Contains "New Media"?	Contains "Internet"?	Contains "e-Commerce"?	Distance to *Test Object*
Training Set	Resume1	Interview	1	1	1	1.00
	Resume2	Interview	1	0	1	1.41
	Resume3	Interview	1	0	0	1.00
	Resume4	Reject	0	0	1	1.73
	Resume5	Reject	0	1	0	1.00
Test Object	Resume6	???	1	1	0	

For the responses to the interactive item, there were three possible categories: *knower*, *misconceiver*, or *mistaker*. Before implementing a prediction algorithm, a set of characteristics defining each response was determined. Several combinations of characteristics were considered. One simple combination, for example, was the x- and y-coordinate location of each line segment. However, in determining whether a response is a *knower*, a *misconceiver*, or a *mistaker*, it is not the *location* of the line segments that is critical, but rather the *relative location* of a line segment to the other line segments, i.e., does a single segment combine with other segments to form a rectangle? Several different methods of representing the relative location of segments were explored. The final representation implemented used 27 characteristics to describe each response.

The first group of characteristics described whether or not prototypical rectangles were formed. Using the labels displayed in Figure 5, this group of characteristics described the relative positions of vertical line segments d_1, d_2, d_3, and d_4, and horizontal line segments d_7, d_9, and d_{10} (d_5 was not included since it could not be used on its own to form a top or bottom side of a prototypical rectangle). There are six combinations of vertical line segments that can form the left and right side of a prototypical rectangle, three combinations of horizontal line segments that can form the top and bottom side of a prototypical rectangle, and twelve combinations of vertical and horizontal segments that can meet to form a vertex of a prototypical rectangle. The algorithm generated a value for each of these 21 combinations, based on whether their distance was within an acceptable range of the prototype. For example, in the prototypical rectangle created using d_1, d_2, d_9, and d_{10}, the distance between the top and bottom sides is 23 pixels. If two of the horizontal segments were within 23±5 pixels of

each other, the algorithm would generate a positive value for that combination. This is how 21 of the 27 characteristics were generated. Because the creation of the prototypical rectangle was an integral component of the scoring, an additional 22nd weighted characteristic was included based on these 21 characteristics.

Figure 5: Line Segments in the Interactive Item



Four of the remaining characteristics were calculated similarly to the process described above, only they reflected whether other rectangles, including the tilted rectangle, were formed. Finally, the numeric response describing the number of rectangles entered was also taken into consideration as a weighted characteristic. All of this information was combined so that each response was described by a set of 27 characteristics. The distance between two responses was then defined using a simple distance formula of all 27 characteristics ($\sqrt{((x_1-x_2)^2 + (y_1-y_2)^2+(z_1-z_2)^2...)}$).

As part of the algorithm development process, many different combinations of characteristics, and values of error ranges, weightings, and distance functions were explored and evaluated. The algorithm chosen and described in this paper showed the best performance at prediction; thus, the decisions made about which algorithmic features to use were data-driven. In other words, while each characteristic used by the algorithm is reflective of the cognitive framework underlying the item (e.g., does the student create a prototypical rectangle), the inclusion or exclusion of certain characteristics and the weighting or interaction of characteristics was determined empirically based on predictive accuracy. Further, the explanatory response provided by students was not considered in the automated algorithm for two reasons. First, it was not considered to provide evidence of misconception as defined by the coding rubric. Second, the purpose of the current research is to explore whether an interactive item can be automatically scored, and the researchers did not want to confound this exploration with automatic assessment of text-based responses. Future research could explore the combination of the interactive components with the text-based response.

The algorithm was evaluated for its predicted effectiveness for the 969 responses about which the two human coders agreed. This was done iteratively. First, a single response was removed as the *test object*, and the remaining 968 responses formed the *training set*. A category was predicted for the *test object*, and that predicted category was compared to the human-coded category to see whether there was agreement. Then, a second response was removed as the *test object*. This process was repeated 969 times, so that each of the responses served as the *test object* and the other 968 responses served as the *training set*. The accuracy across all 969 trials was averaged to determine the overall accuracy of the algorithm.

The algorithm was also evaluated for its predictions on the 18 responses about which the human coders disagreed. A response cannot be included in the *training set* if it does not have a single defined category. Therefore, these 18 responses could not be included in the *training set*. Instead, the 969 responses about which the human coders agreed were used for the *training set*, and each of the 18 responses was iteratively used as the *test object*.

Results

The algorithm predicted a code for each answer in 0–16 milliseconds. When coding with three categories, *knower*, *misconceiver*, and *mistaker*, the algorithm correctly predicted the category for 94.6% of all responses (Cohen's κ=0.903). Table 3 shows the accuracy of the algorithm in terms of the number of responses in each actual and predicted category, both in raw numbers and percentages. The cells along the diagonal, in bold, represent correct predictions, and sum to 94.6%.

Table 3: Accuracy of the Prediction Algorithm on the 969 Double-Coded Responses

		Predicted Code		
		Knower	***Misconceiver***	***Mistaker***
Actual Code	*Knower*	**562 (58%)**	4 (0.4%)	6 (0.6%)
	Misconceiver	2 (0.2%)	**140 (14.4%)**	6 (0.6%)
	Mistaker	24 (2.5%)	10 (1.0%)	**215 (22.2%)**

Instead of coding responses into three categories, it is also possible to code into only two categories. One method is to code all responses as *correct* or *incorrect*, where all *mistaker* and *misconceiver* responses are re-coded as *incorrect*, and all *knower* responses are re-coded as *correct*. This type of coding would be useful for a more traditional achievement assessment, rather than a diagnostic measure. When responses were re-coded in this way, the algorithm performed slightly better than when predicting with three categories, and correctly predicted the category for 96.3% of all responses (Cohen's κ=0.927). A second method is to code all responses as *misconceiver* or *non-misconceiver*, where all *mistaker* and *knower* responses are re-coded as *non-misconceiver*. This type of coding would be useful for an assessment that was only returning diagnostic information, and was not returning any traditional achievement feedback. When responses were re-coded in this way, the algorithm correctly predicted the category for 97.7% of all responses (Cohen's κ=0.914).

Prediction Errors

There were errors made when predicting a category for the 969 responses about which the two human coders agreed. This sub-section describes these errors that were made in predicting one of the three categories (*knower*, *misconceiver*, and *mistaker*), as providing this type of diagnostic feedback is the intent of the overall DGA project. While the accuracy rate for this prediction was very high, it is important to consider the ramifications of the improperly predicted codes. The ultimate goal of an automated scoring algorithm is to implement the algorithm in a real-time classroom testing environment where teachers can be provided with instant scores. Therefore, it is important to explore the responses that were incorrectly predicted, as these errors could have real-world consequences, namely: 1) a student operating under the misconception who is not identified as such will not receive additional instruction to help overcome the misconception; or 2) a student who has a thorough understanding of the concept but is incorrectly identified as having the misconception will receive superfluous instruction and perhaps miss more advanced instruction during this time.

Table 4 shows the types of errors made by the algorithm. Thirty-one percent of these errors (*mistaker* predicted as *misconceiver* and *misconceiver* predicted as *mistaker*) are considered less significant, because the actual and predicted codes both represent an *incorrect* answer, and because, from a diagnostic standpoint, a student would not be classified as a *misconceiver* based on the result of a single item. These types of errors do not incur the two real-world consequences defined above. The remaining 70% of the errors, however, are significant, as they represent a *correct* response being scored as *incorrect*, or vice versa.

Table 4: Types of Errors Made by the Prediction Algorithm

		Predicted Code		
		Knower	***Misconceiver***	***Mistaker***
Actual Code	*Knower*	—	4 (8%)	6 (12%)
	Misconceiver	2 (4%)	—	6 (12%)
	Mistaker	24 (46%)	10 (19%)	—

There were 24 *mistaker* responses that were predicted as *knower*. For the majority of these responses (15 responses), the picture created represented the knowledge of a *knower*, but the number of rectangles entered by the student did not match the picture; Figure 6(a) (next page) is an example of this type of response. For these errors, the weighting of the relative location of the line segments over-powered the number reported by students during prediction, thus resulting in a prediction of *knower*. For four out of the 24 *mistaker* responses predicted as *knower*, the answers were coded by the researchers as *incorrect* because one of the rectangles was missing a side, as displayed in Figure 6(b) (next page). For these errors, it is likely that the relative locations of the line segments were close enough to forming rectangles that the algorithm made a *knower* prediction. The remaining five responses of this type were other assorted incorrect responses that were predicted to be correct; Figures 6(c) and (d) (page 25) are two examples of these responses. For these errors, the relative location of the line segments made it appear to the algorithm that prototypical and tilted rectangles were being created, although they were, in fact, not.

Figure 6: Sample *Mistaker* Responses Predicted as *Knower*

Figure 6(a):

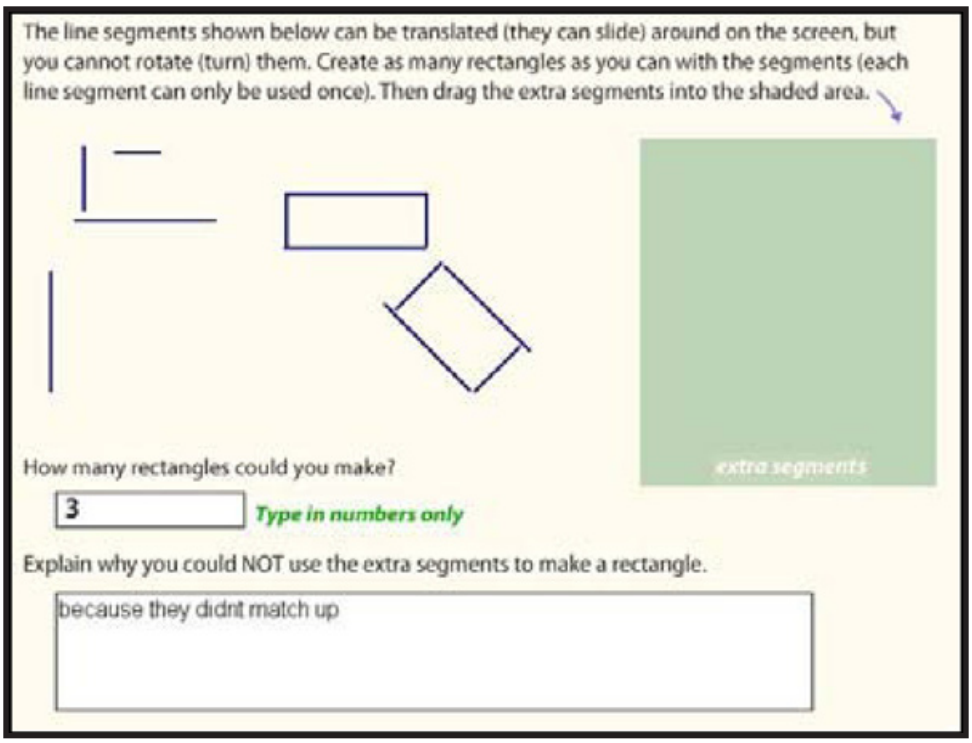


Figure 6(b):

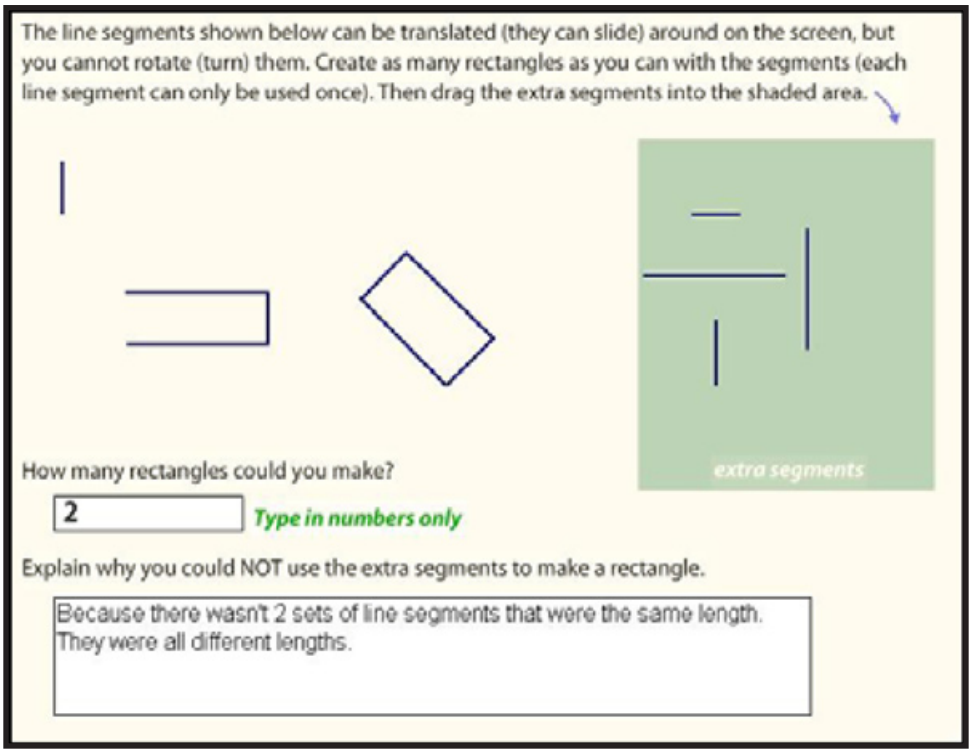


Figure 6(c):

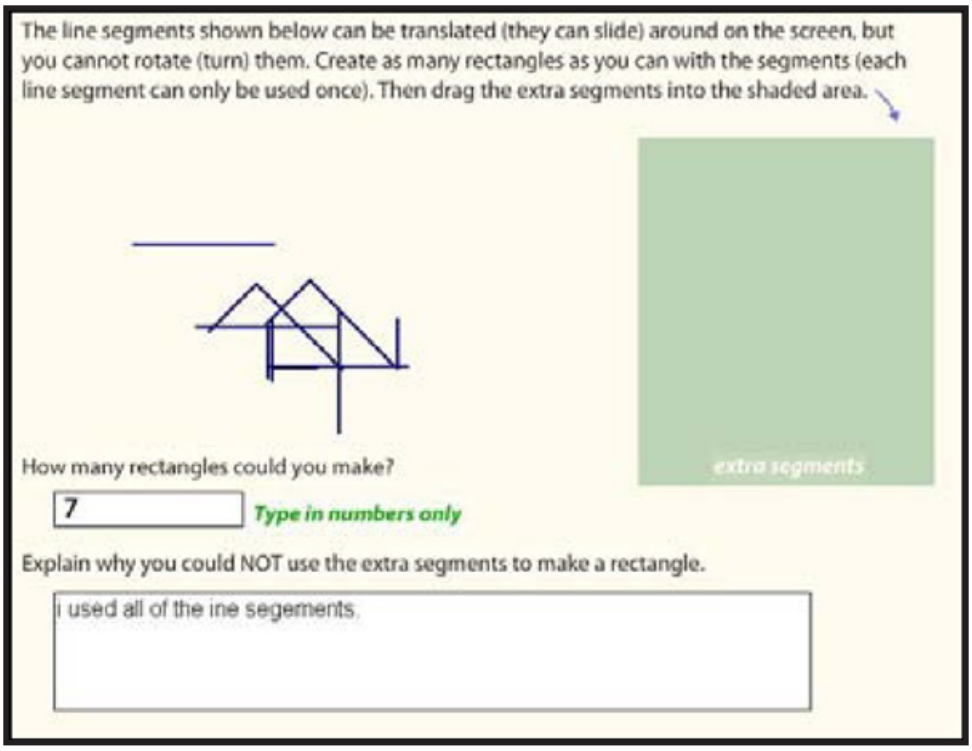
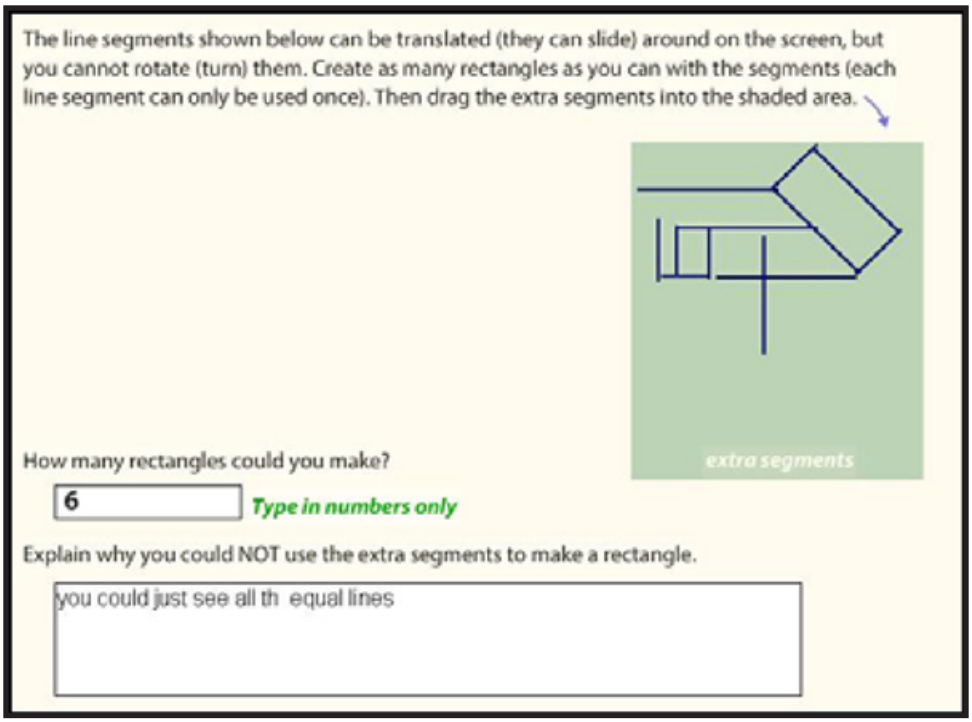
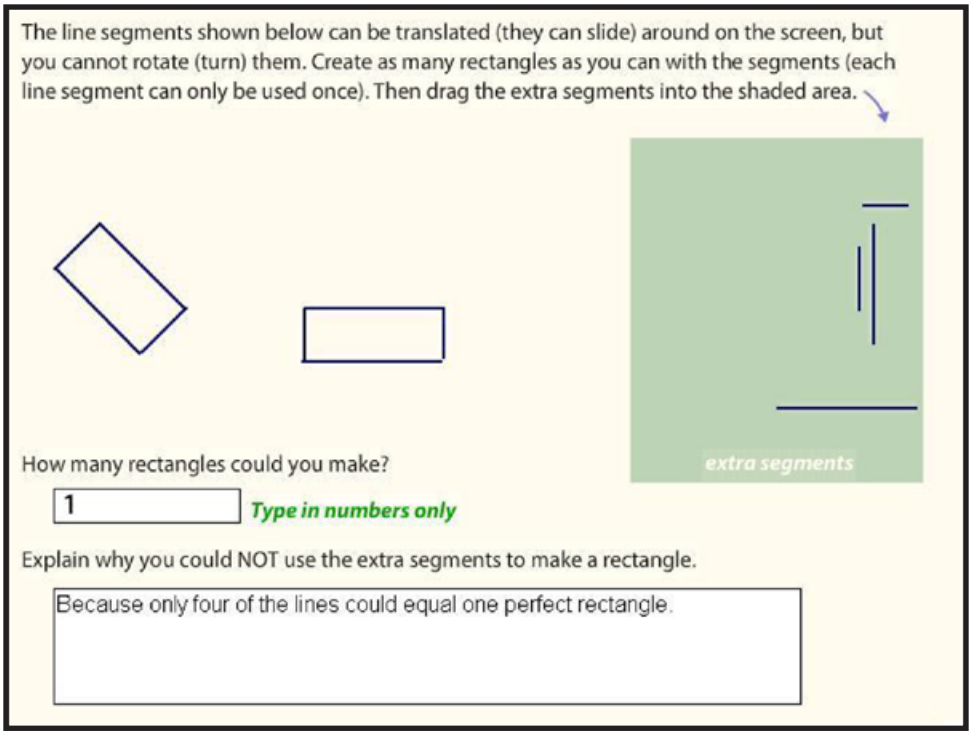


Figure 6(d):



There were two *misconceiver* responses predicted as *knower*. For both of these responses, the picture created represented the knowledge of a *knower*, but the student reported only one rectangle was created; Figure 7 is an example of one of these responses. In these errors, as in the previous case, the weighting of the location of the line segments over-powered the number reported by students during prediction.

Figure 7: Sample *Misconceiver* Response Predicted as *Knower*



The final two categories of errors includes the ten *knower* responses that were predicted to be *incorrect* (four predicted as *misconceiver* and six predicted as *mistaker*). For all of these responses, the picture created was unique and non-similar to other responses; Figure 8 (next page) is an example of one of the responses predicted to be a *misconceiver*, and Figure 9 (next page) is an example of one of the responses predicted to be a *mistaker*. It is likely that these unusual responses were not similar to other responses in their *neighborhood*, thus resulting in unreliable and incorrect predictions. Automating accurate predictions of unusual objects is a common challenge in prediction systems.

Figure 8: Sample *Knower* Response Predicted as *Misconceiver*

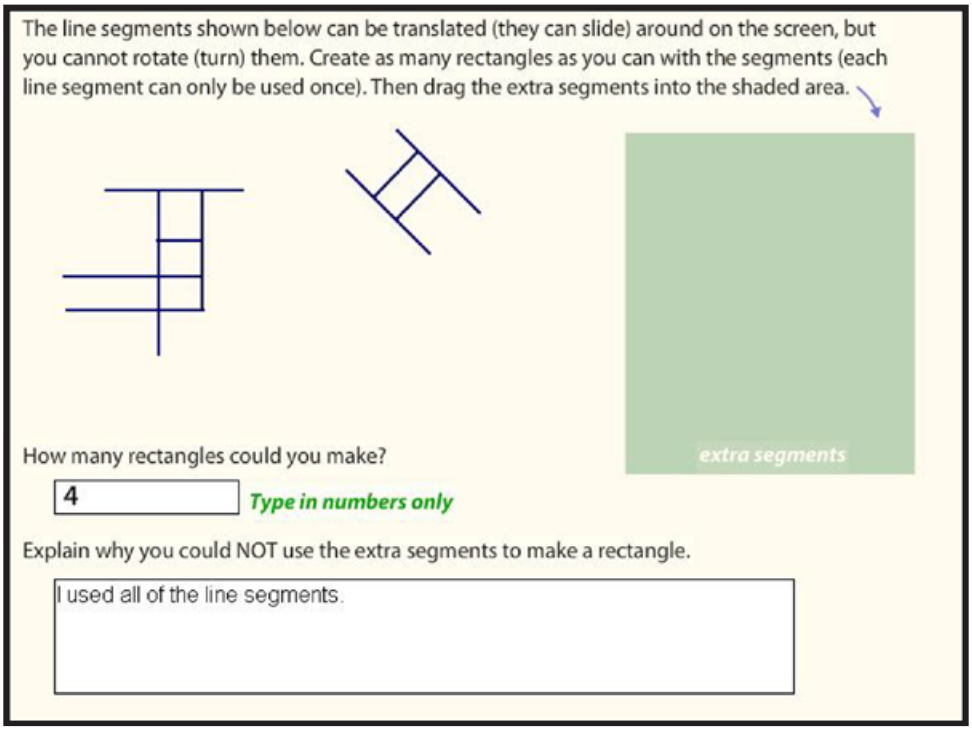
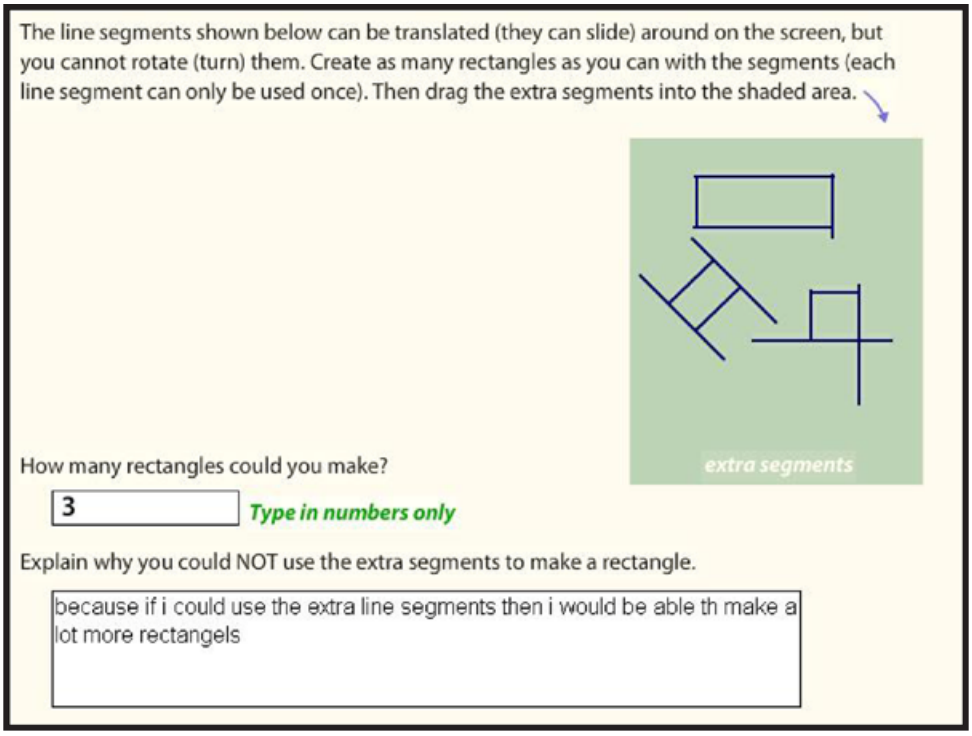


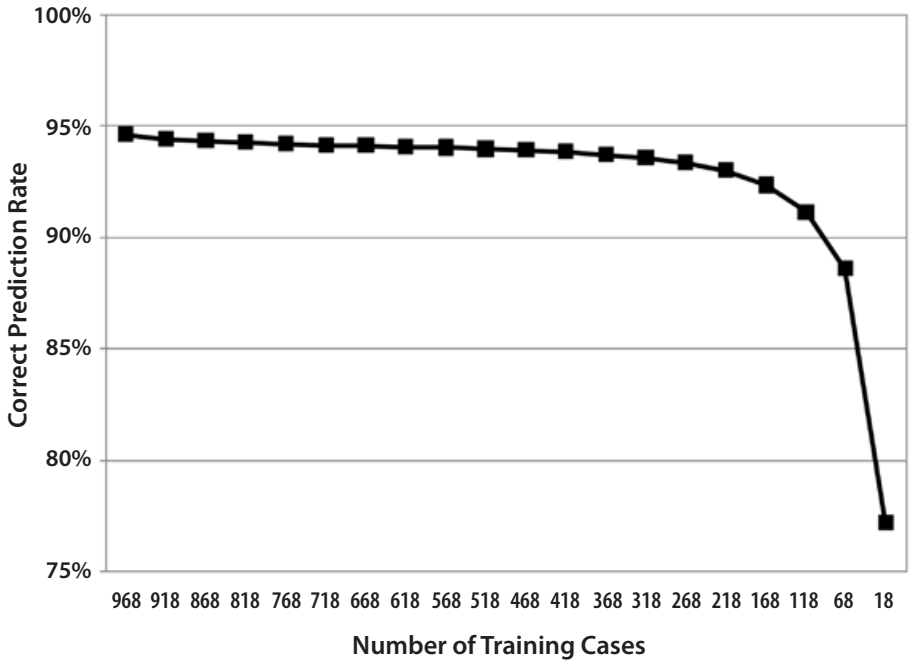
Figure 9: Sample *Knower* Response Predicted as *Mistaker*



Size of the Training Set

It might seem inevitable to obtain high predictive accuracy with upwards of 1,000 responses on which to base a prediction. Therefore, the accuracy of the algorithm was tested using a smaller set of responses for prediction. To begin, a random sample of 50 responses was removed from the *training set*, leaving 918 responses. Then, a predicted code was made for each of the 969 responses, based only on the codes for this reduced set of 918 responses. Because the removal of the 50 responses was random, it was repeated for 1,000 trials, and the average accuracy was calculated based on all trials. The average accuracy was 94.5%. Next, 100 responses were removed from the *training set*, leaving 868 coded responses on which to base predictions. A code was then predicted for each of the 969 responses based on this reduced *training set*, and this trial was repeated 1,000 times. This process was repeated until there were only 18 responses remaining in the *training set*. Figure 10 shows the average accuracy rates for each trial. The accuracy remained above 90% with as few as 118 cases on which to base predictions, after which the accuracy decreased.

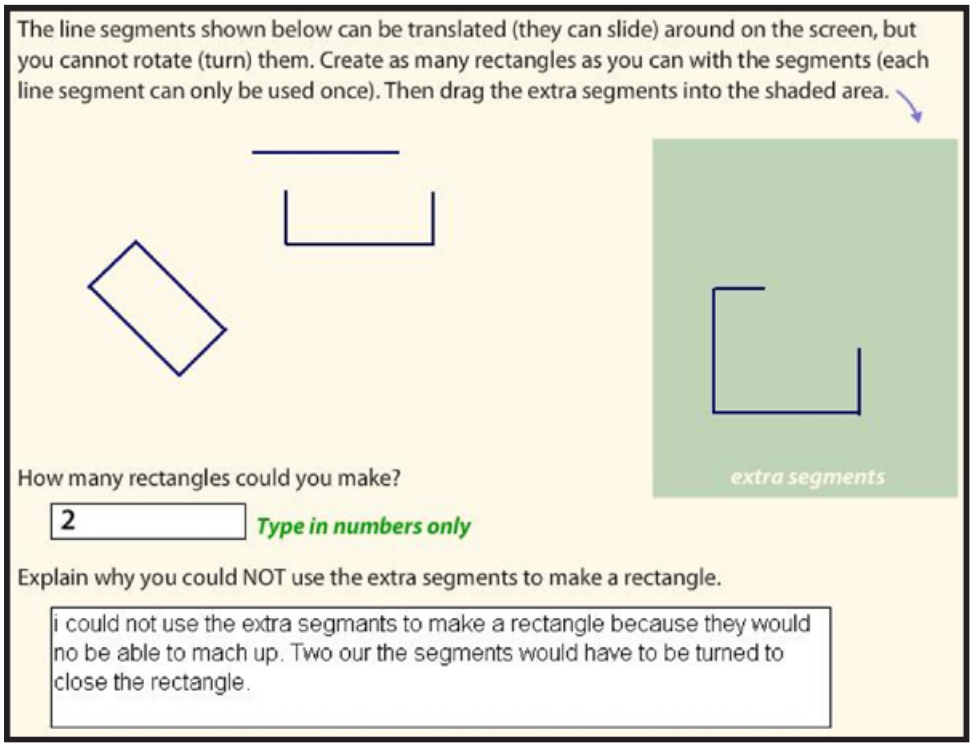
Figure 10: Average Accuracy based on Varying Number of Cases



Human-Coding Disagreements

For 17 of the 18 responses about which there was human disagreement, the category predicted by the algorithm agreed with one of the two human coders (94.4%). Two of the 18 responses were blank, but were incorrectly coded by one human coder as *mistaker*. For one of the 18 responses, displayed in Figure 11, one researcher coded the response as a *knower*, allowing for the incomplete rectangle to satisfy the requirement of a prototypical rectangle, and the other researcher coded the response as a *mistaker*, judging that the rectangle was not complete enough to satisfy the requirement. This was a subjective decision. The algorithm predicted that this response was a *mistaker*.

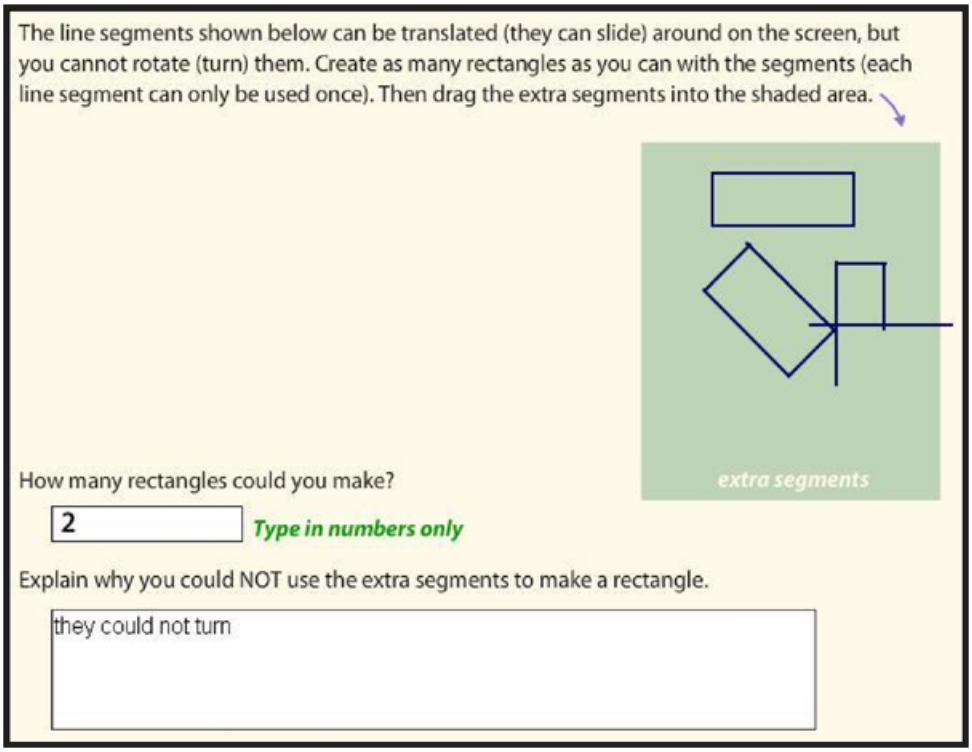
Figure 11: Response for which Human Coders Disagreed based on Subjective Interpretation



For the remaining 15 responses, upon review of the human scoring, the disagreements were attributed to human error in interpreting the rubric, and using a strict interpretation of the coding rubric, one of the two human codes was assigned to each response as a consensus code. When comparing the predicted code to this newly assigned consensus code, the algorithm showed 53% agreement; there were seven responses where the predicted code did not match the consensus code. Two of these responses

were *misconceiver* responses that the algorithm predicted as *mistaker* (an "insignificant" error as described earlier). For the remaining five responses, a sample of which is displayed in Figure 12, the consensus code was either a *misconceiver* or a *mistaker*, and the predicted code was *knower*. For these responses, similar to one type of error described earlier, the location of the line segments did, in fact, represent a *knower* response, but the number of reported rectangles did not. For these errors, the weighting of the relative location of the line segments over-powered the number reported by students in the prediction process.

Figure 12: Sample Response for which Human Consensus Code Disagreed with Predicted Code



Discussion

Formative assessment is a critical component of classroom practice. For assessment results to truly inform instruction, they must provide timely feedback to teachers. While traditional multiple-choice assessments can provide instant results, they are of limited depth. Open-response items provide richer information, but instant scoring is often untenable. Research efforts have explored effective methods of automatically scoring open-response items, both in essay scoring and in mathematics.

The research presented here expands this base of research into the area of geometry in the middle grades. Responses to an interactive item designed to assess misconceptions related to Shape Properties were hand-coded. An algorithm was developed to automatically predict a code for the student responses. The algorithm based predictions on the relative locations of the line segments within the item. The algorithm showed high accuracy when compared to human coding. This accuracy was high when coding responses both dichotomously and diagnostically. Ultimately, the goal of creating an automated coding system would be to allow interactive items to be included in a larger diagnostic assessment that returned instant feedback to teachers. This research provides initial evidence that this is possible.

There were errors in prediction. Seventy-percent of the errors were considered "significant," i.e., an incorrect answer was predicted as correct, or vice versa. However, upon closer inspection, for some errors, specifically those displayed in Figures 6(a) and (b) (page 24), it is possible that the algorithm's prediction would be closer to a teacher's scoring. Because the coding rubric used by the human coders was designed for research purposes, and would be used independently by two separate coders, it strictly defined *knower* responses as those where the number of rectangles created matched the number reported (excluding the one leniency described earlier). However, a classroom teacher might take more leeway in the grading of these types of responses than did the researchers and, in fact, code them as correct, as was predicted by the algorithm. Additionally, while researchers were allowed to use subjective judgment when segments did not line up perfectly, they should not have accepted rectangles that were completely missing a side. Similar to the last type of error, however, these types of responses might be more loosely interpreted by a classroom teacher as correct. Thus, in a classroom setting, some of the errors that are considered significant might be more closely aligned to a teacher's interpretation than to the researchers' coding, and thus the errors might have less impact than originally assumed. It should be noted that these are the researchers' conjectures about the practical significance of some of the predictive errors, and are not based on classroom evaluations.

Finally, there is evidence that in future trials of automatic scoring algorithms, a smaller *training set* can be used while still maintaining predictive accuracy. While it is premature to set a specific acceptable cut point in predictive accuracy, this research provides evidence that the algorithm had an accuracy rate of greater than 90% with a greatly reduced number of cases on which to base predictions. Therefore, future trials exploring the automatic coding of similar interactive items could be based on a smaller collected sample of responses. This finding is encouraging given the efforts required to recruit teachers to participate in exploratory research. Further, if this item were to be implemented in a classroom setting, a limited number of cases in the *training set* could be provided along with the testing system while still maintaining high predictive accuracy.

There are two main contributions of this research. First, it serves as initial evidence that traditional algorithms from artificial intelligence can be used to accurately and automatically code student responses, both dichotomously and diagnostically, to an interactive geometry item for the middle grades. This finding opens up great potential for the development of future interactive items that can offer K–12 mathematics teachers the opportunity to gain deeper insight into their students' thinking *and* be automatically scored to provide teachers with instant feedback on student understanding. The majority of research in automated scoring has focused on essay scoring. Research exploring the automatic scoring of mathematics items has focused on rational expressions and graphs, typically at the undergraduate or graduate level. This research expands that base into the area of geometry in the middle grades. Teachers of K–12 mathematics have often been constrained to traditional multiple choice items if they want an assessment that can be instantly scored. This research is a first step towards leveraging existing scoring technology for the mathematics domain in the middle grades.

Second, this research provides evidence that innovative and interactive item types may not necessarily need to be hand-scored. While this paper focused on a single item, future research will be aimed at developing a *class* of similar items, all of which could be scored with the same underlying algorithm. For example, a set of items could ask students to build parallelograms, rectangles, and squares, and provide line segments which could create both prototypical and non-prototypical shapes. The same automatic scoring algorithm could be applied to all of these items; item developers would only have to specify the defining characteristics of correct and misconception responses. Other item sets might explore different types of interactivity, but utilize the same underlying algorithm described here. By providing evidence that innovative and interactive items can be automatically scored, this research supports the idea these types of items should be developed and researched.

There are two main limitations of this work. First, it is based on a single item, rather than a class of items. However, this work was *intended* as a proof-of-concept. Future work will involve developing additional interactive items and exploring the accuracy of the algorithm on those items. Second, although the accuracy was very high, there remained a group of responses that were incorrectly scored, some with significant errors. While no algorithm will have perfect accuracy, just as two human coders would not have perfect agreement, predictive errors should not be discounted. One way to combat this reality is to develop a measure of confidence to accompany predicted scores. In this way, the algorithm could *flag* a group of responses about which it was less certain about its accuracy. It might base this certainty on the distance of the *test object* to its nearest neighbor, for example, or on a separate set of characteristics altogether. The group of flagged responses, which would contain both responses that could be accurately scored and responses that could not, would be removed for hand-coding by the classroom teacher. Future work will focus on exploring such a confidence rating.

Endnotes

1. Due to a technical error in the administration of the teacher survey, ethnicity information was not collected about participating teachers.
2. Regions defined by the United States census: http://www.census.gov/geo/www/us_regdiv.pdf.
3. Locales defined by the National Center for Education Statistics Common Core of Data: http://nces.ed.gov/ccd/.
4. The explanatory section of the item was included for the original purpose of item administration (using student responses to develop closed-response options), and was not considered relevant to classifying an individual response as a *knower*.
5. As stated earlier, a student would never be classified as a *misconceiver* based on his or her response to a single item.

References

Airasian, P.W. (1991). *Classroom Assessment*. New York, NY: McGraw-Hill.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3).

Bennett, R.E., Morley M., & Quardt, D. (2000a). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13*, 303–322.

Bennett, R.E., Morley M., & Quardt, D. (2000b). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*(4), 294–309.

Bennett, R.E. & Sebrechts, M.M. (1994). *The Accuracy of Automatic Qualitative Analyses of Constructed-Response Solutions to Algebra Word Problems* (GE Board Professional Report No. 91-03P). Princeton, NJ: Educational Testing Service.

Bennett, R.E., Steffen, M., Singley, M.K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*(2), 162–176.

Ben-Simon, A. & Bennett, R.E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6*(1).

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.

Breese, J.S., Heckerman, D., & Kadie, C. (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering* (Technical Report MSR-TR-98-12). Microsoft Research, Microsoft Corporation.

Burstein, J. & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing* (pp. 68–75). College Park, MD.

Carroll, W.M. (1998). Geometric knowledge of middle school students in a reform-based mathematics curriculum. *School Science and Mathematics, 98*, 188–197.

Chan, K. & Yeung, D. (2000a). An efficient syntactic approach to structural analysis of on-line handwritten mathematical expressions. *Pattern Recognition, 33*, 375–384.

Chan, K. & Yeung, D. (2000b). Mathematical expression recognition: A survey. *International Journal on Document Analysis and Recognition, 3*(1), 3–15.

Clements, D.H. (2003). Teaching and learning geometry. In J. Kilpatrick, W.G. Martin, & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 151–178). Reston, VA: NCTM.

Clements, D.H. & Battista, M.T. (1992). Geometric and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics* (pp. 420–464). New York, NY: Macmillan.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1).

Hershkowitz, R. (1989). Visualization in geometry: two sides of the coin. *Focus on Learning Problems in Mathematics, 11*, 61–76.

Hiebert, J., & Carpenter, T.P. (1992). Learning and teaching with understanding. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 65–97). New York, NY: MacMillan.

Mestre, J. (1987). Why should mathematics and science teachers be interested in cognitive research findings? *Academic Connections* (pp. 3–5, 8–11). New York, NY: The College Board.

Nichols, P. (2005). *Evidence for Interpretation and Use of Scores From an Automated Essay Scorer* (PEM Research Report 05-02). Pearson Educational Measurement.

Page, E.B. (2003). Project Essay Grade: PEG. In M.D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Page, E.B. & Petersen, N.S. (1995). The computer moves into essay grading. *Phi Delta Kappan, 76*, 561–565.

Popham, W.J. (1995). *Classroom Assessment: What Teachers Need to Know*. Needham Heights, MA: Allyn and Baco.

Popham, W.J. (October 2006). *Defining and enhancing formative assessment*. Paper presented at the Annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, CA.

Popham, W.J. (2008). *Transformative Assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Resnick, L. (1983). Mathematics and science learning: A new conception. *Science, 220*, 477–478.

Resnick, L., Nesher, P., Leonard, F., Magone, M., Omamson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20*(1), 8–27.

Rudner, L.M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetric Essay Scoring System. *Journal of Technology, Learning, and Assessment, 4*(4).

Rudner, L.M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2).

Rudner, L.M., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation, 7*(26).

Russell, M., O'Dwyer, L.M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*, 414–424.

Sangwin, C. (2004). Assessing mathematics automatically using computer algebra and the Internet. *Teaching Mathematics and Its Applications: An International Journal of the IMA, 23*(1), 1–14.

Sebrechts, M.M., Bennett, R.E., & Rock, D.A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856–862.

Shermis, M. & Burstein, J. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Singley, M.K. & Bennett, R.E. (1998). *Validation and Extension of the Mathematical Expression Response Type: Applications of Schema Theory to Automatic Scoring and Item Generation in Mathematics* (GRE Board Report No. 93-24P). Educational Testing Services.

Smith, J.P., diSessa, A.A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. The *Journal of the Learning Sciences, 3*, 115–163.

Wang, J. & Brown, M.S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2).

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 319–330.

Vinner, S., & Hershkowitz, R. (1980). Concept images and common cognitive paths in the development of some simple geometrical concepts. In R. Karplis (Ed.), *Proceedings of the Fourth International Conference for the Psychology of Mathematics Education* (pp. 177–184), Berkeley, CA: University of California, Lawrence Hall of Science.

Author Note

This research was supported by the Institute of Education Sciences, award number R305A080231. The Diagnostic Geometry Assessment project, which included the development of the coding rubric and the coding of open-response items, is a collaboration amongst researchers at the Technology and Assessment Center at Boston College and the Center for Leadership and Learning Communities at the Education Development Center, Inc.

Author Biography

Jessica Masters is a Senior Research Associate at Boston College's Technology and Assessment Study Collaborative. She received her Ph.D. in Computer Science, with a focus on educational technology, from the University of California at Santa Cruz. Her research focuses on the intersections of technology, education, and assessment. She can be reached via e-mail at jessica.masters@bc.edu.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College